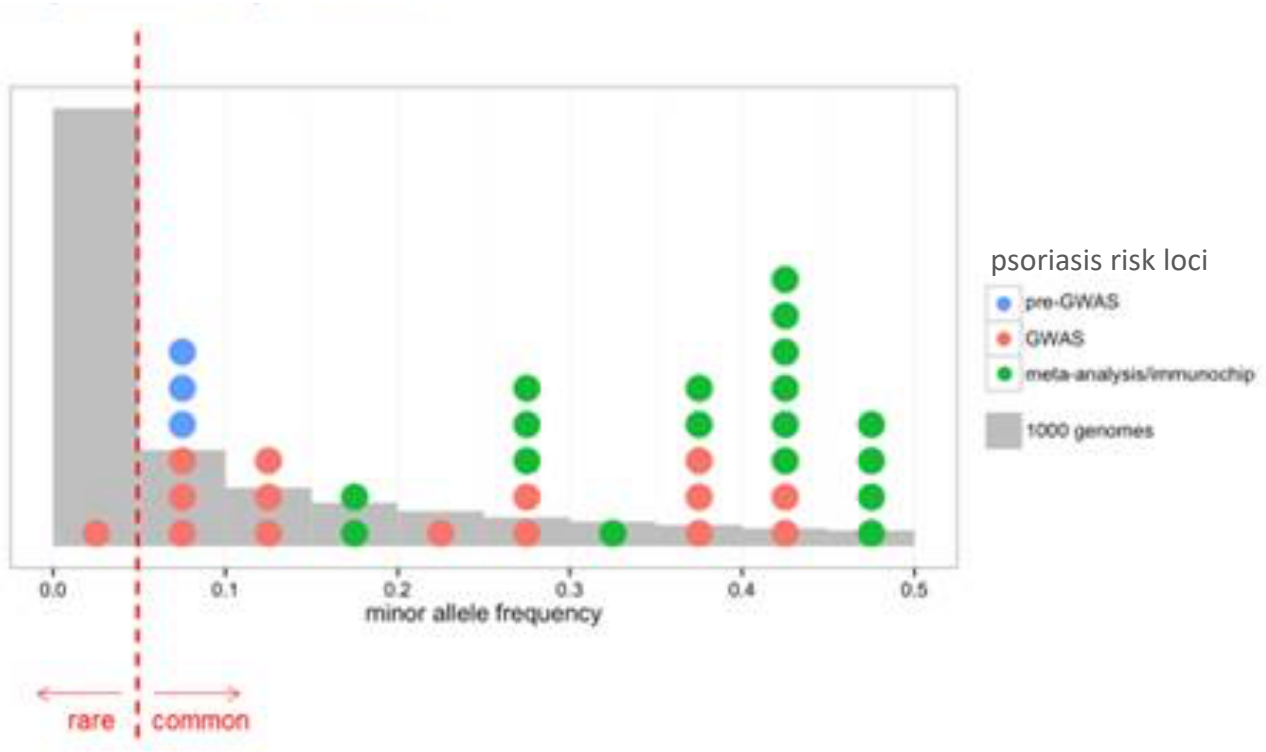
Rare Variant Association Testing

Why are we interested in rare variants?

Most genetic variants are rare

The contribution of rare variation is not well captured by GWAS



Why are we interested in rare variants?

We can now access this rare variation cost effectively

Move towards a more complete the understanding of the genetic architecture of each trait and disease

Identify new risk loci

Rare protein coding variation useful to ‘pinpoint’ causal genes at established loci

To identify genes where loss-of-function variation is protective, which may be new targets for therapy

Why do we expect rare variants to contribute to common disease?

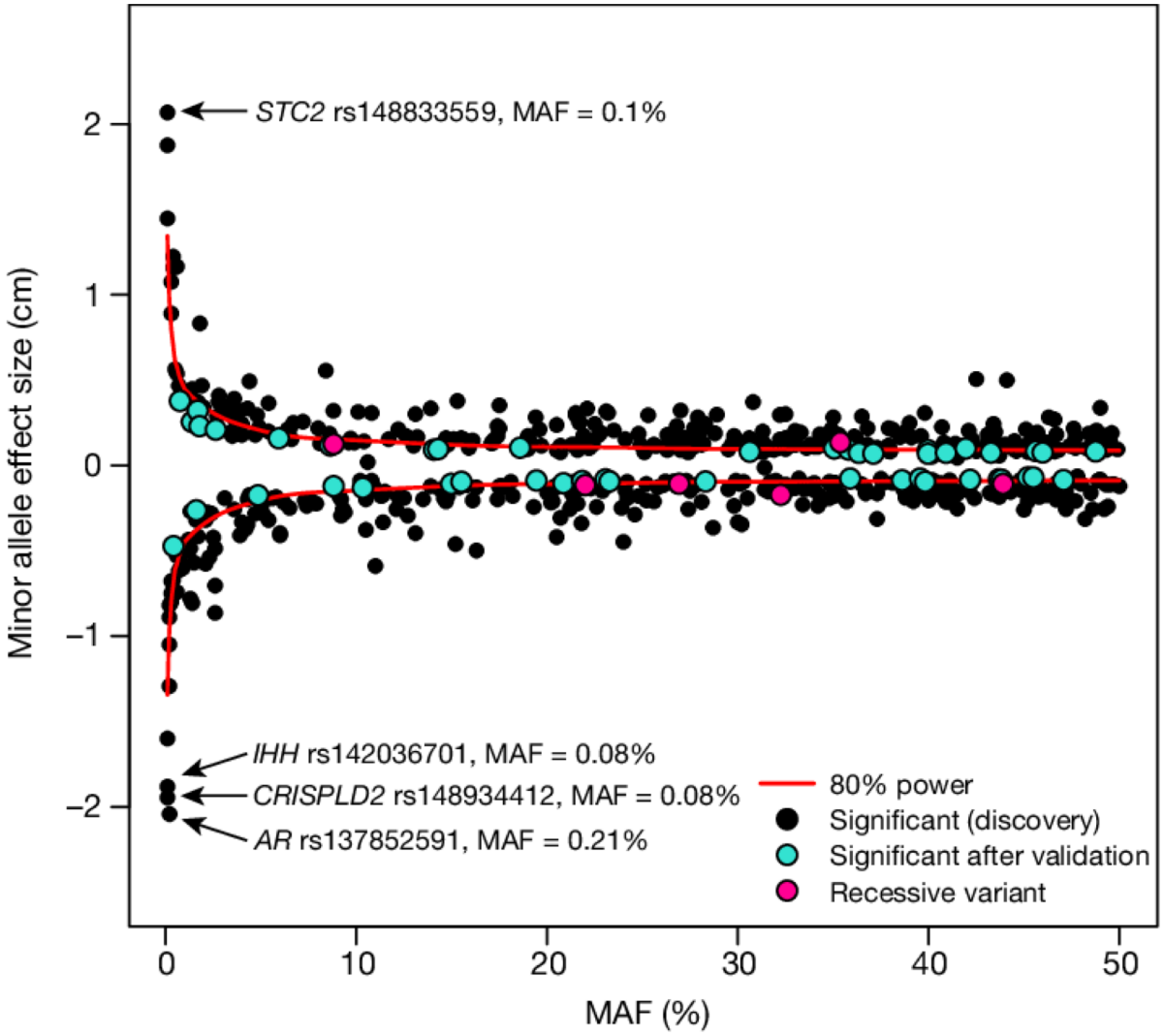
Evolutionary theory indicates that deleterious alleles are likely to be rare

Rare variants are known to play a role in human disease

- Mendelian disorders
- rare forms of common disease

Emerging empirical evidence of the role of rare variants in common diseases/traits

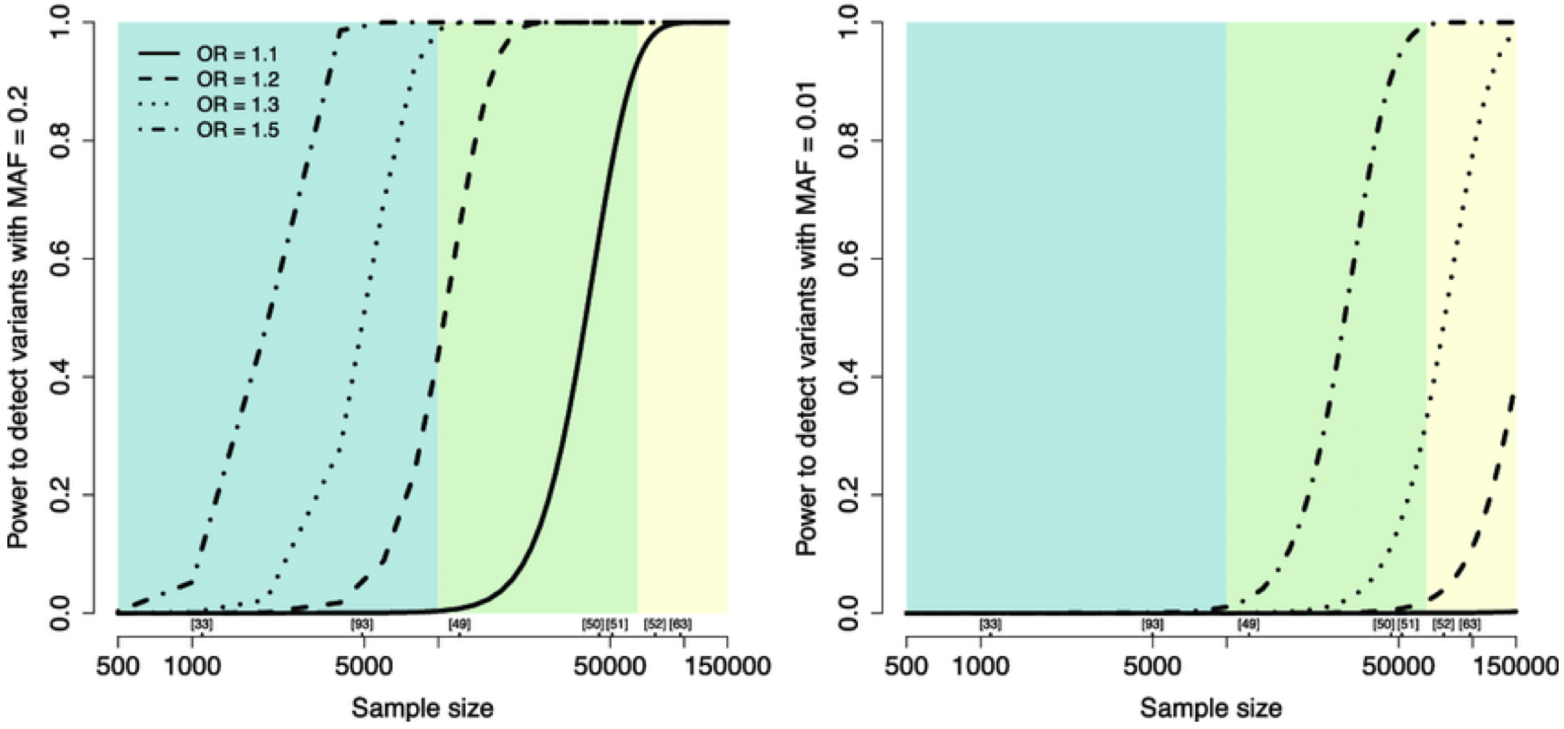
Rare variants and height



n=700k

Marouli et al Nature 2017

Single variant association testing



The rarity of the alleles impacts on statistical power to identify phenotypic effects

Studies of individual rare variants will be underpowered unless sample sizes or effect sizes are very large

de Lange et al 2015

How do we overcome these limitations?

Rather than test individual variants for association, we can consider groups of variants with similar functional effects

For example all variants with MAF < 0.01 in the protein coding region of the same gene

Count or score presence or absence of rare variants per individual and use this variant score to predict trait values using standard regression models

If all variants are causal and they have the same direction of effect, this leads to large increase in power

Toy example

Variant		Individuals												
1. A:T		0/0	0/0	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
2. AA:A		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/1	0/0
3. T:C		0/1	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
4. T:C		0/0	0/1	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
5. A:G		0/0	0/0	1/1	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
6. T:G		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/1	0/1	0/1
7. G:C		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
8. G:T		0/0	0/0	0/1	0/0	0/0	0/0	…	0/0	0/1	0/0	0/0	0/0	0/0
9. A:AT		0/0	0/0	0/0	0/0	0/1	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
10. A:T		0/1	0/0	0/0	0/0	0/0	0/0	…	0/1	0/0	0/0	0/0	0/0	0/0

Toy example

Count up the number of rare variants per individual

Variant		Individuals												
1. A:T		0/0	0/0	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
2. AA:A		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/1	0/0
3. T:C		0/1	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
4. T:C		0/0	0/1	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
5. A:G		0/0	0/0	1/1	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
6. T:G		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/1	0/1	0/1
7. G:C		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
8. G:T		0/0	0/0	0/1	0/0	0/0	0/0	…	0/0	0/1	0/0	0/0	0/0	0/0
9. A:AT		0/0	0/0	0/0	0/0	0/1	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
10. A:T		0/1	0/0	0/0	0/0	0/0	0/0	…	0/1	0/0	0/0	0/0	0/0	0/0
Variant score		2	2	3	3	1	1		1	1	1	1	2	1

Use this score as the predictor in a regression model

Morris and Zeggini 2010

Toy example

Collapse to a binary presence/absence of any rare variant - cohort allelic sum test (CAST)

Variant		Individuals												
1. A:T		0/0	0/0	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
2. AA:A		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/1	0/0
3. T:C		0/1	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
4. T:C		0/0	0/1	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
5. A:G		0/0	0/0	1/1	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
6. T:G		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/1	0/1	0/1
7. G:C		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
8. G:T		0/0	0/0	0/1	0/0	0/0	0/0	…	0/0	0/1	0/0	0/0	0/0	0/0
9. A:AT		0/0	0/0	0/0	0/0	0/1	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
10. A:T		0/1	0/0	0/0	0/0	0/0	0/0	…	0/1	0/0	0/0	0/0	0/0	0/0
Variant score		1	1	1	1	1	0		1	1	0	1	1	1

Use this score as the predictor in a regression model

Morgenthaler and Thilly 2007

Burden tests - limitations

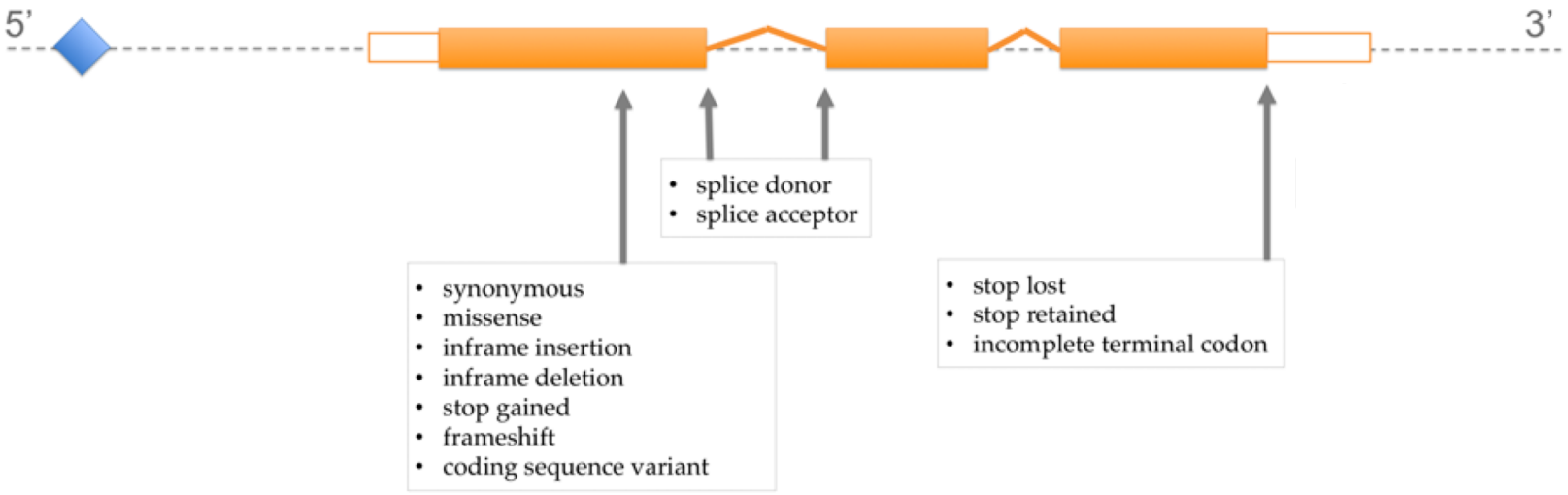
However, including all observed rare variants within a gene would likely include many **neutral** variants and potentially variants with **opposite direction** of effect – which will impact power to detect a true association

Two options

1. Try to select/weight variants that have similar functional consequences in an attempt to focus on variants with the same magnitude and direction of effect

2. Use a statistical framework that is robust to neutral and opposing directions of effect

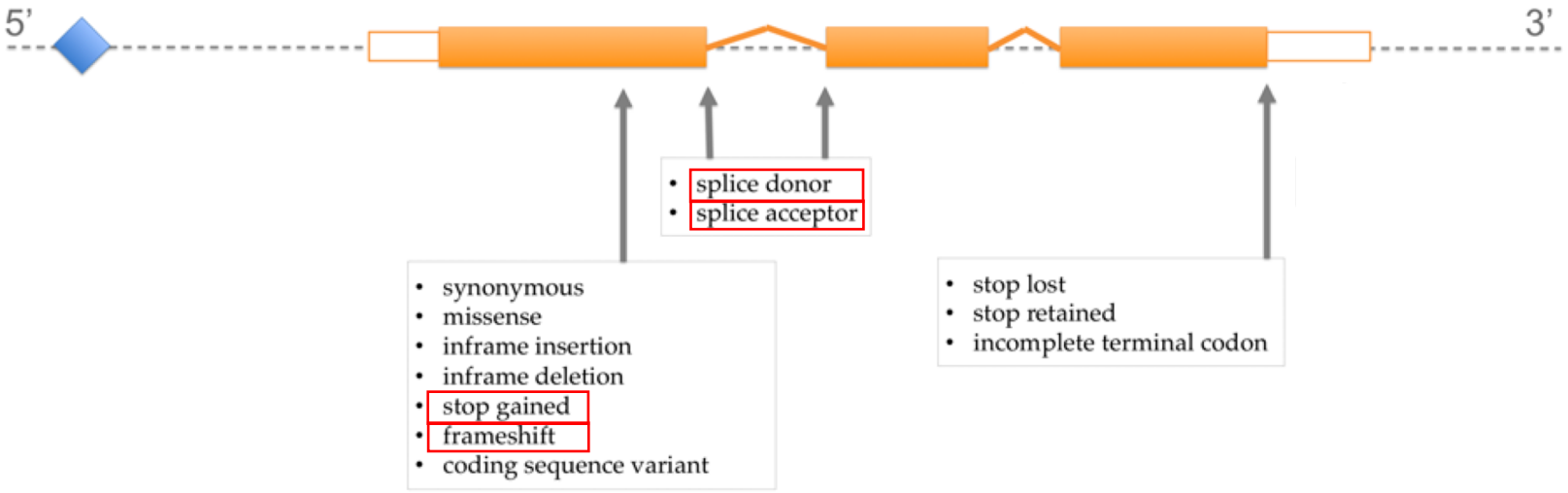
Consequences of protein coding variation



Loss-of-function variation

One class of variants where functional prediction is simpler are those which truncate the resulting protein product

stop-gain, *frameshift, splice disrupting*



Loss-of-function variation

Protein truncating variation may be subject to nonsense-mediated decay (NMD), a cellular mechanism that prevents the expression of truncated proteins

To a first approximation, PTVs are thus likely to result in the same functional consequence - loss of function (LoF)

Therefore LoF variants are a set of variants that are naturally combined in a burden test

Caution – not all PTVs are subject to NMD (~final 5% of gene) and this class of variants are enriched for false positive variant calls (LOFTEE can help)

Toy example

Variant		Individuals												
1. A:T		0/0	0/0	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
2. AA:A		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/1	0/0
3. T:C		0/1	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
4. T:C		0/0	0/1	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
5. A:G		0/0	0/0	1/1	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
6. T:G		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/1	0/1	0/1
7. G:C		0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
8. G:T		0/0	0/0	0/1	0/0	0/0	0/0	…	0/0	0/1	0/0	0/0	0/0	0/0
9. A:AT		0/0	0/0	0/0	0/0	0/1	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
10. A:T		0/1	0/0	0/0	0/0	0/0	0/0	…	0/1	0/0	0/0	0/0	0/0	0/0

Toy example

Variant	consequence	Individuals												
1. A:T	synonymous	0/0	0/0	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
2. AA:A	frameshift	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/1	0/0
3. T:C	stop gain	0/1	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
4. T:C	missense	0/0	0/1	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
5. A:G	stop gain	0/0	0/0	1/1	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
6. T:G	missense	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/1	0/1	0/1
7. G:C	missense	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
8. G:T	splice site	0/0	0/0	0/1	0/0	0/0	0/0	…	0/0	0/1	0/0	0/0	0/0	0/0
9. A:AT	frameshift	0/0	0/0	0/0	0/0	0/1	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
10. A:T	missense	0/1	0/0	0/0	0/0	0/0	0/0	…	0/1	0/0	0/0	0/0	0/0	0/0

Toy example

Variant	consequence	Individuals												
1. A:T	synonymous	0/0	0/0	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
2. AA:A	frameshift	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/1	0/0
3. T:C	stop gain	0/1	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
4. T:C	missense	0/0	0/1	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
5. A:G	stop gain	0/0	0/0	1/1	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
6. T:G	missense	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/1	0/1	0/1
7. G:C	missense	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
8. G:T	splice site	0/0	0/0	0/1	0/0	0/0	0/0	…	0/0	0/1	0/0	0/0	0/0	0/0
9. A:AT	frameshift	0/0	0/0	0/0	0/0	0/1	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
10. A:T	missense	0/1	0/0	0/0	0/0	0/0	0/0	…	0/1	0/0	0/0	0/0	0/0	0/0

Toy example

Variant	consequence	Individuals												
1. A:T	synonymous	0/0	0/0	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
2. AA:A	frameshift	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/1	0/0
3. T:C	stop gain	0/1	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
4. T:C	missense	0/0	0/1	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
5. A:G	stop gain	0/0	0/0	1/1	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
6. T:G	missense	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/1	0/1	0/1
7. G:C	missense	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
8. G:T	splice site	0/0	0/0	0/1	0/0	0/0	0/0	…	0/0	0/1	0/0	0/0	0/0	0/0
9. A:AT	frameshift	0/0	0/0	0/0	0/0	0/1	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
10. A:T	missense	0/1	0/0	0/0	0/0	0/0	0/0	…	0/1	0/0	0/0	0/0	0/0	0/0
Variant score		1	0	3	1	1	0		0	1	0	0	1	0

Toy example

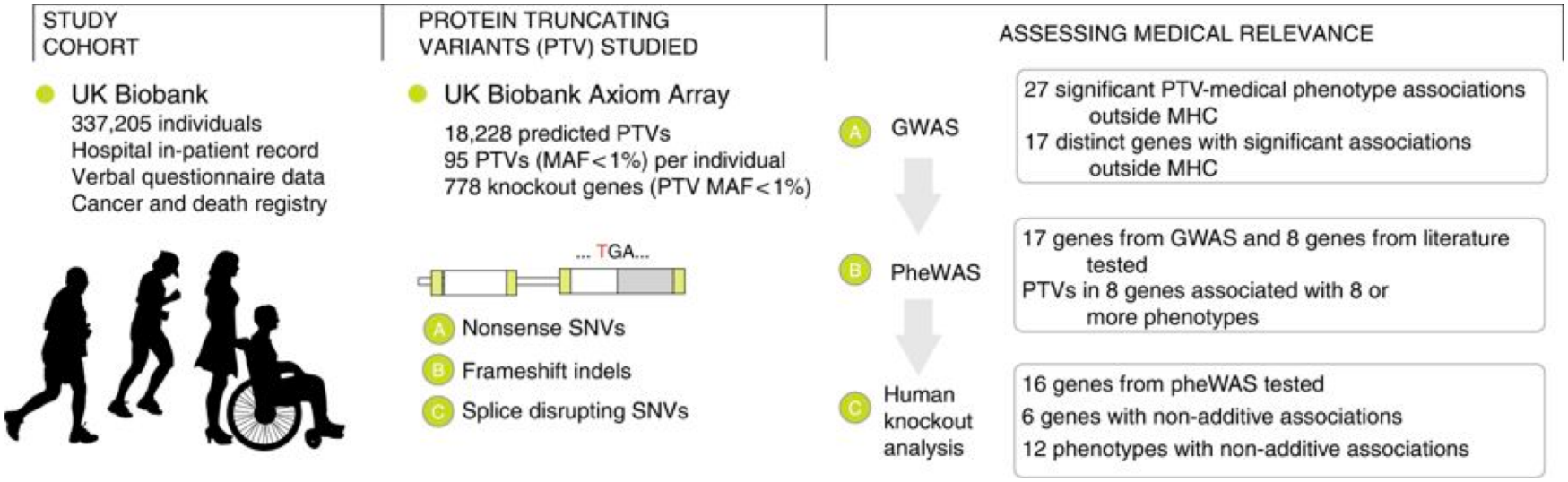
Variant	consequence	Individuals												
1. A:T	synonymous	0/0	0/0	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
2. AA:A	frameshift	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/1	0/0
3. T:C	stop gain	0/1	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
4. T:C	missense	0/0	0/1	0/0	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
5. A:G	stop gain	0/0	0/0	1/1	0/1	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
6. T:G	missense	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/1	0/1	0/1
7. G:C	missense	0/0	0/0	0/0	0/0	0/0	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
8. G:T	splice site	0/0	0/0	0/1	0/0	0/0	0/0	…	0/0	0/1	0/0	0/0	0/0	0/0
9. A:AT	frameshift	0/0	0/0	0/0	0/0	0/1	0/0	…	0/0	0/0	0/0	0/0	0/0	0/0
10. A:T	missense	0/1	0/0	0/0	0/0	0/0	0/0	…	0/1	0/0	0/0	0/0	0/0	0/0
Variant score		1	0	2	1	1	0		0	1	0	0	1	0

Use this score as the predictor in a regression model

LoF variation in UKBB

Exome sequencing data on 50k individuals released next week… remaining 450k to follow

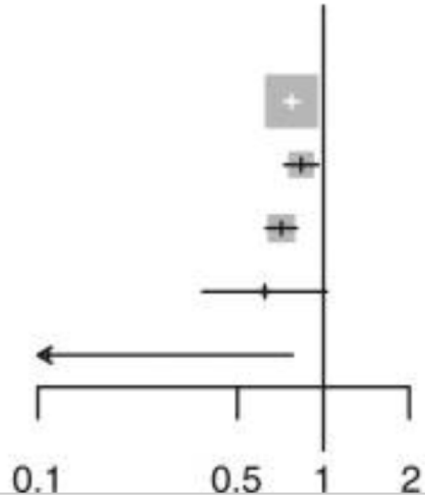
Some insight from 18k genotyped rare LoF variants



DeBoever et al 2018

Protective LoF variants in IFIH1 in immune related disorders

Outcome	Cases	Controls	OR	95% CI	*p*-value
Hypothyroidism	23591	381978	0.77	[0.73; 0.82]	4.7×10^{-16}
Hyperthyrodism	4477	401092	0.84	[0.73; 0.96]	0.0096
Psoriasis	5920	399649	0.71	[0.63; 0.81]	8.8×10^{-8}
Type 1 diabetes	130	405439	0.62	[0.38; 1.03]	0.063
Vitiligo	344	405225	0.11	[0.02; 0.78]	0.027



Burden test of four rare LoF variants in IFIH1

Emdin et al 2018

Burden tests - summary

Burden tests assume that all the rare variants in a region are causal and affect the phenotype in the same direction with similar magnitudes

They suffer from a substantial loss of power when these assumptions are violated

Restrict to specific classes of variation (ie LoF)

Weighting schemes have also been described to weight variants based on predicted functional consequence, conservation, biochemical properties or use frequency as a proxy for functional consequence

An alternative approach

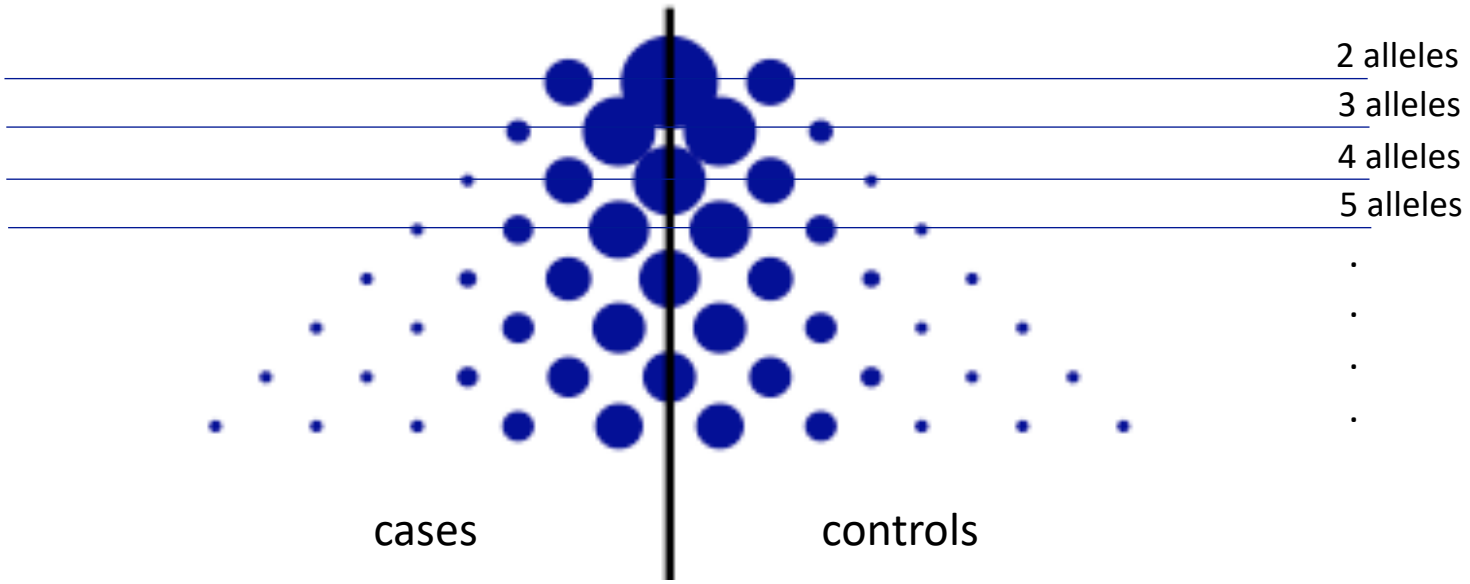
We expect for many genes that there will be a mixture of variants with effects on phenotype and no effects on phenotype

For some genes we expect to see variants that have effects in opposite directions

Can we test whether a mixture of neutral, risk, and protective variants are present in a gene?

C-alpha test – Neale et al 2011 an approach for testing for the presence of this mixture of effects across a set of rare variants

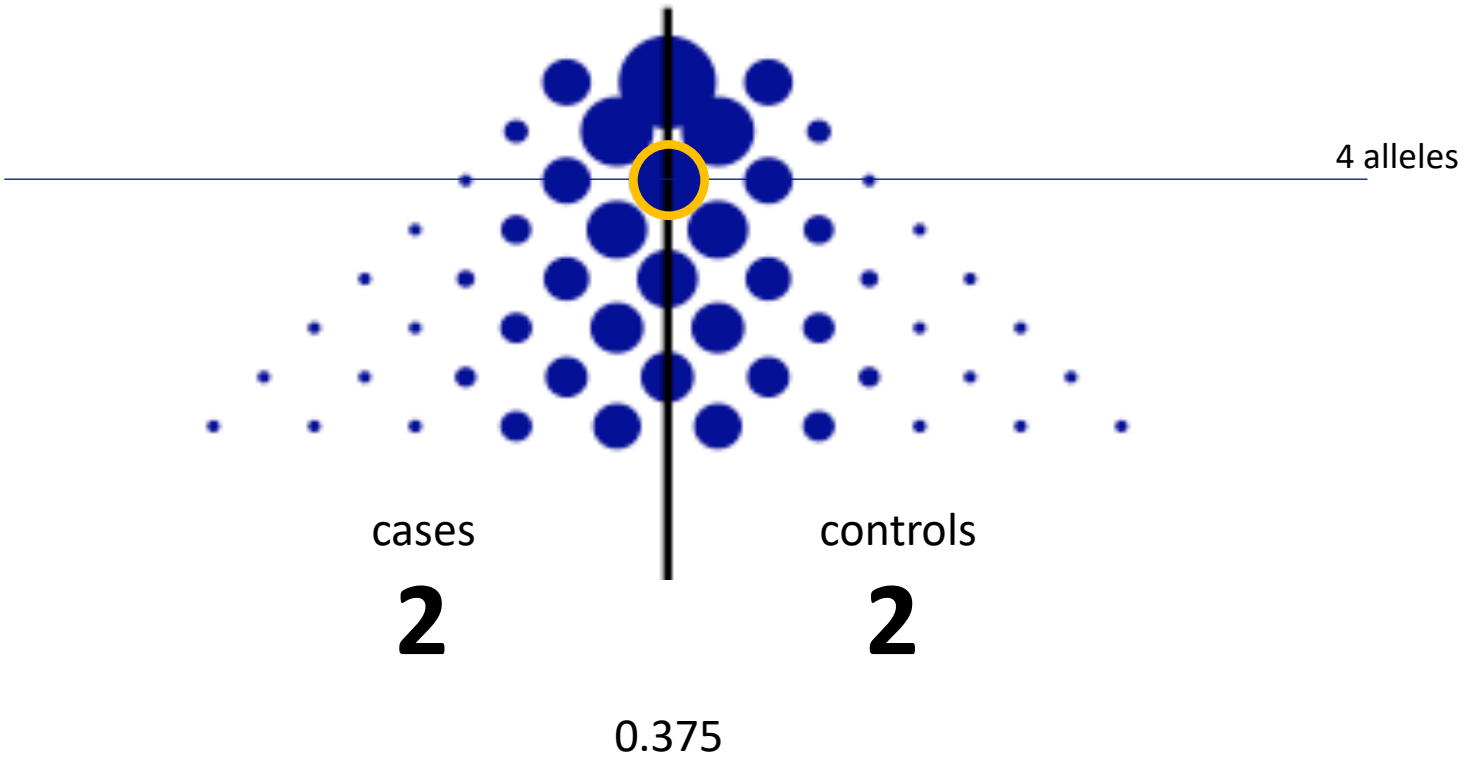
Binomial Expectation



Each row shows a different number of copies of the rare variant [2-9]
We align the variation by counting # in cases vs. # controls

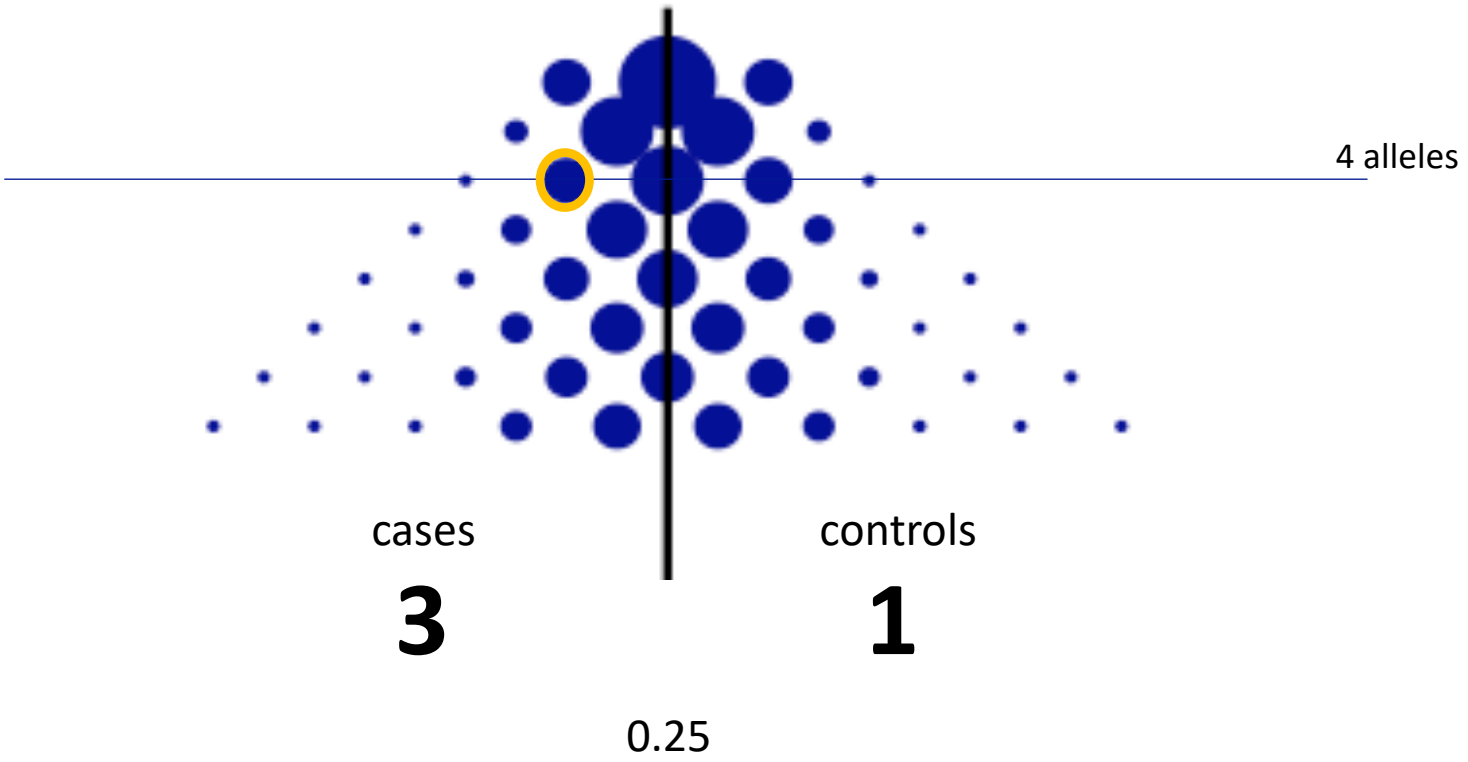
Neale et al 2011

Binomial Expectation



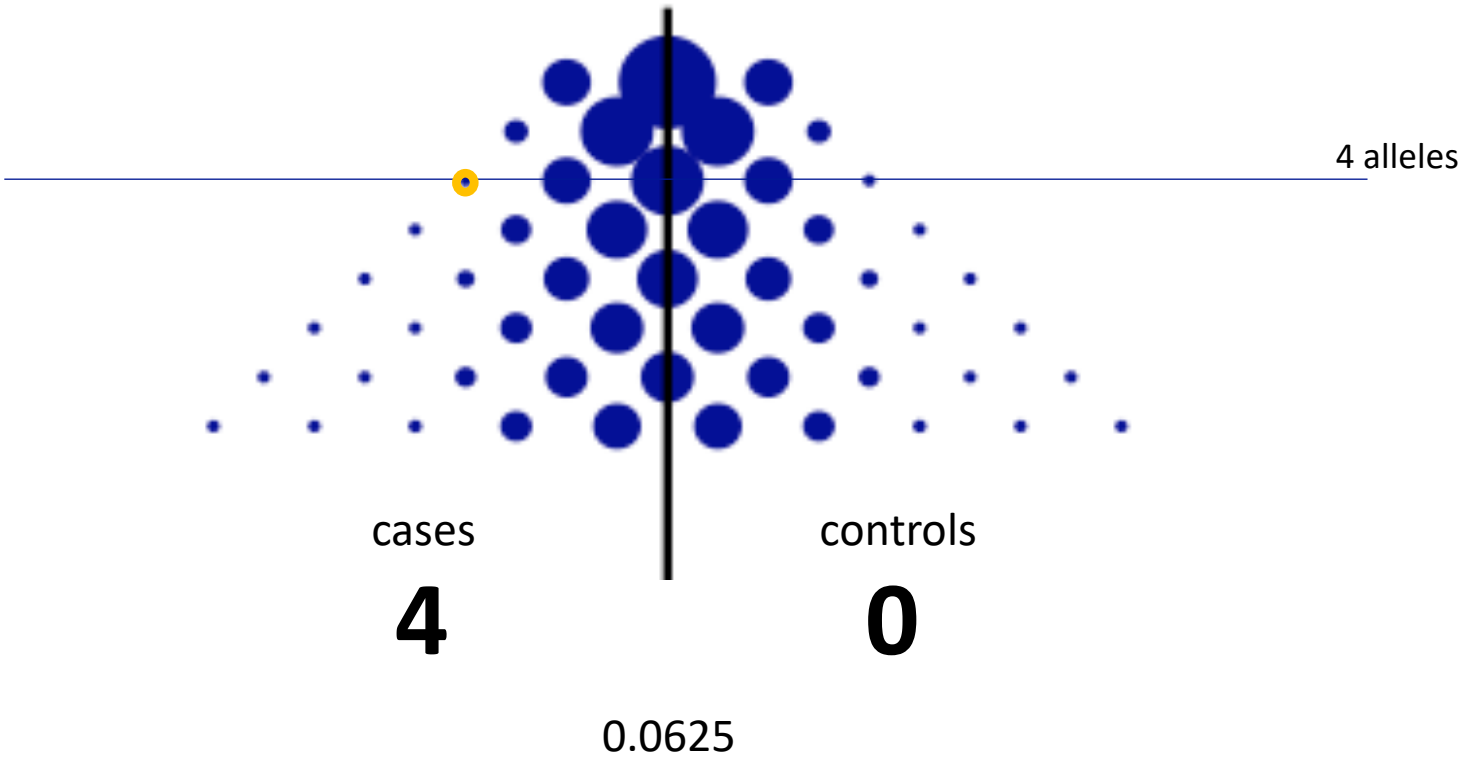
Neale et al 2011

Binomial Expectation



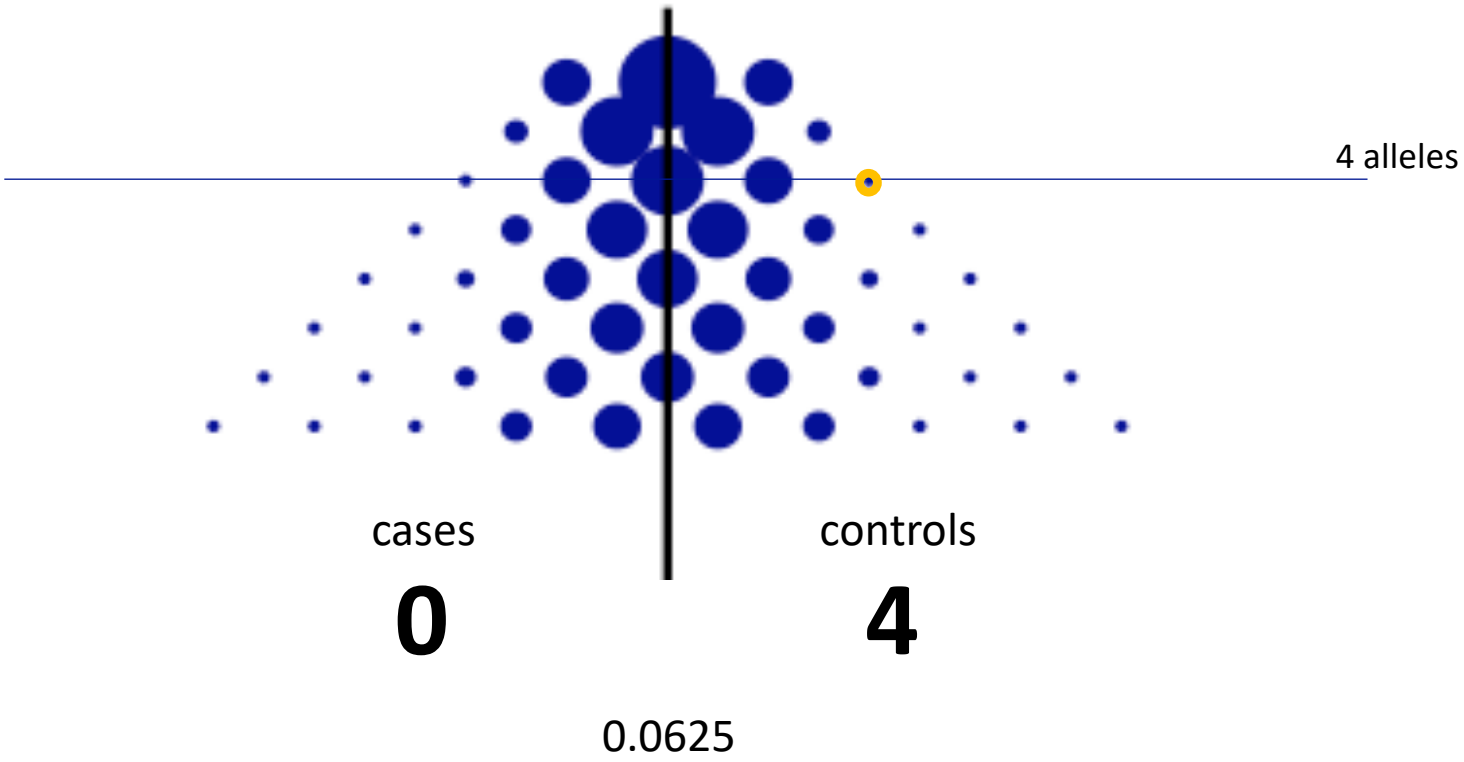
Neale et al 2011

Binomial Expectation



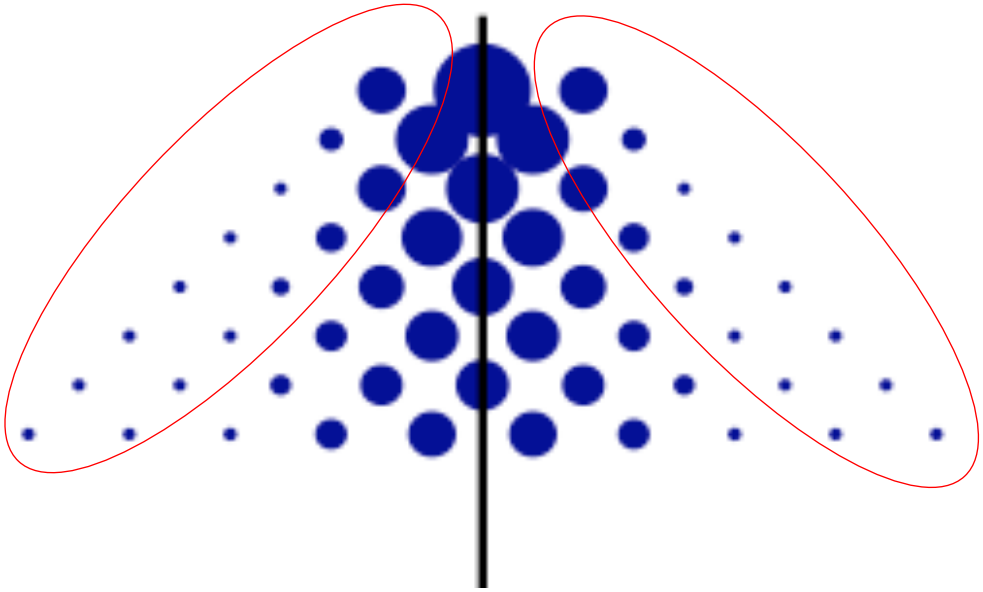
Neale et al 2011

Binomial Expectation



Neale et al 2011

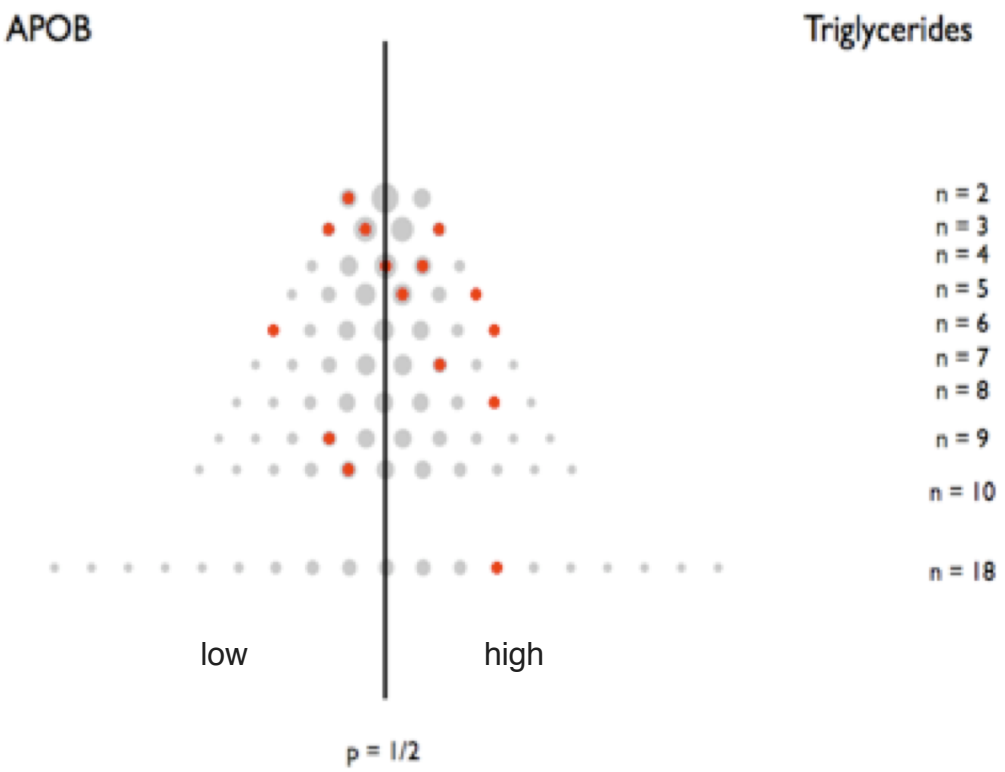
What does signal look like?



overdispersion, or increase in the binomial variance in the allele distributions

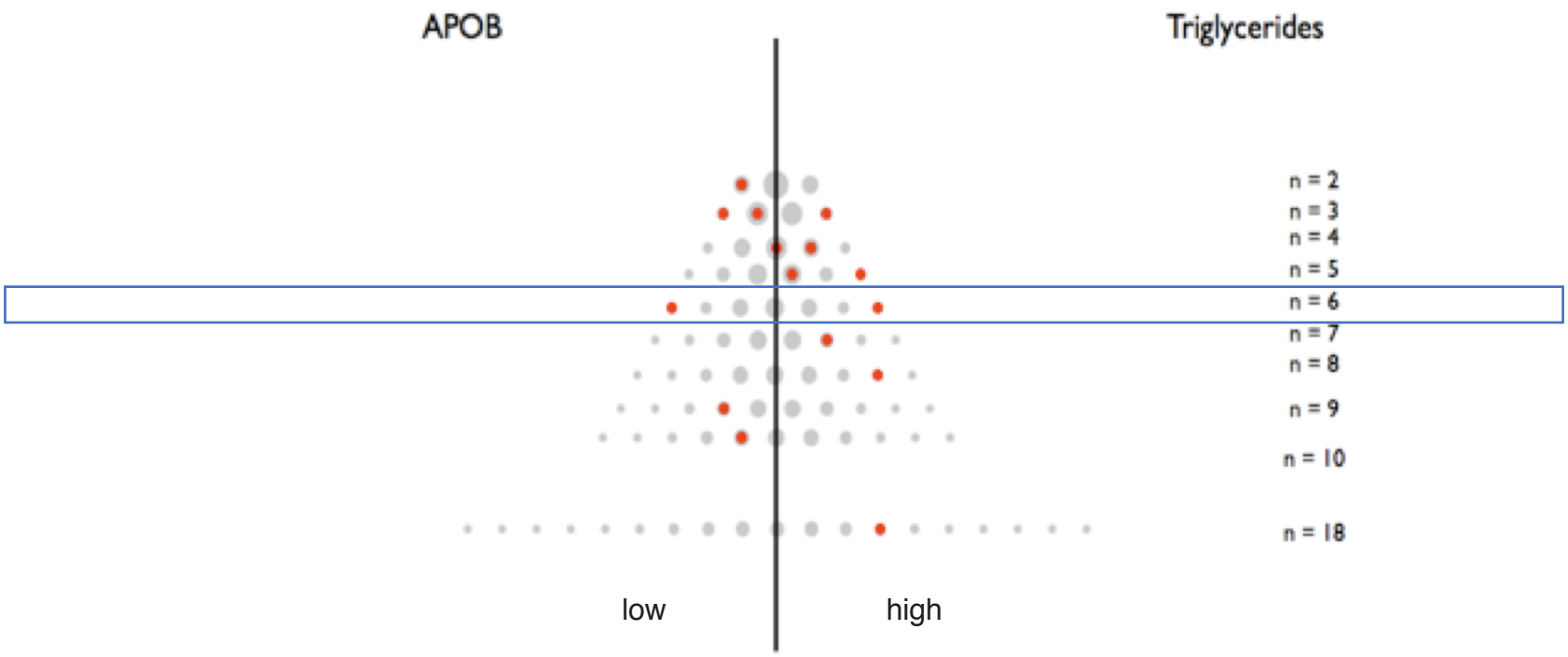
Neale et al 2011

APOB – Triglycerides



Distribution of allele counts of 15 rare variants observed in 100 individuals who have high triglycerides and 100 individuals who have low triglycerides

APOB – Triglycerides



We observe two variants with six alleles

One has 6 copies in cases and 0 in controls

One has 6 copies in controls and 0 in cases

C-alpha and SKAT

SNP-set (Sequence) Kernel Association Test

Can be considered as a generalized C-alpha test

- does not require permutation but calculates the p value analytically
- allows for covariate adjustment
- accommodates either dichotomous or continuous phenotypes

Range of extensions to SKAT, including SKAT-O optimal linear combination of SKAT and burden test

Summary

Sequencing data can now be generated at a scale that we can start to systematically examine the role of rare variation in population based studies of common complex disease

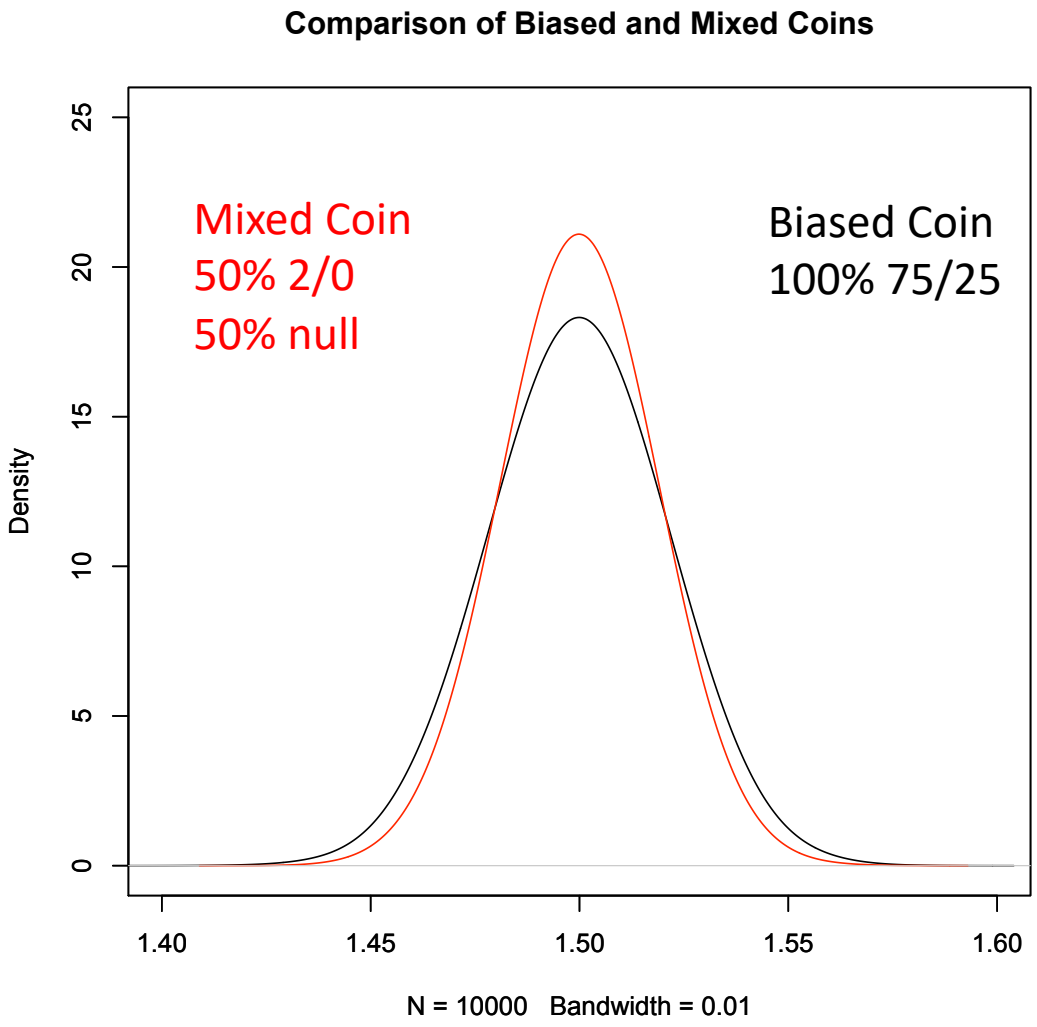
However, power is limited to detect association of individual variants, unless very large effect sizes or very large sample sizes

Combining variants within functional units (ie genes) offers scope to overcome the power limitations

Simple burden tests make a strong assumption that all rare variants in a set are causal and associated with a trait with the same direction and magnitude of effect, but this is appropriate when considering LoF variation (practical this afternoon)

Variance component tests (C-alpha and SKAT) are robust to groups of variants that include variants with positive effects, negative effects and those that are neutral

It’s a mixture we’re looking for



Coin	2/0	1/1	0/2
Mixed Coin	0.625	0.25	0.125
Biased Coin	0.5625	0.375	0.0625